



## StrandSpecRNAseqEval Documentation

**Description:** Evaluates strand-specific yeast (*S. cerevisiae*) RNA-seq libraries.  
**Author:** Moran Yassour, [gp-help@broadinstitute.org](mailto:gp-help@broadinstitute.org)

### Summary

Many methods have been developed for strand-specific RNA-seq that do not lose the information about which strand was transcribed, as standard RNA-seq methods do. This enhances the value of an RNA-seq experiment, particularly in densely-coded microbial genomes.

For evaluation of new *Saccharomyces cerevisiae* libraries, we have developed this module to analyze:

- the complexity of the library, meaning the number of distinct, unique read start positions
- strand specificity of the reads, by comparing the mapped reads to the expected transcribed strand based on known annotations
- evenness and continuity of transcript coverage
- coverage at the 5' and 3' ends
- performance in digital expression profiling relative to reference expression measurements

The methods implemented in this module and the results to which some users will want to compare their results are described fully in Levin et al (2010). The supplementary website for this paper is located [here](#).

**Note: This module is for use with *S. cerevisiae* libraries only.**

### Usage/Example

This module uses a SAM file as input. SAM files tend to be very large. The Broad public server limits uploads to a maximum of 2 GB. If your SAM file is larger than this, you will need to [download GenePattern and install it on your local system](#) (and your local system must run Linux), and run the module there. For more information on the SAM format, see the specification: <http://samtools.sourceforge.net/SAM-1.3.pdf>.

The module outputs one Excel file, tab-delimited files for expression data, genome browser tracks, and a number of figures. The Excel file contains the numerical result of each evaluation step, and the PNG figures accompany the results.

This module, using the Example Data listed below, takes approximately 10 to 12 hours to run on the Broad public server.

## References

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*. 2010; doi:10.1038/nmeth.1491.

## Parameters

Name	Description
sam.file (required)	The input SAM file. For more information on the SAM format, see the specification: <a href="http://samtools.sourceforge.net/SAM-1.3.pdf">http://samtools.sourceforge.net/SAM-1.3.pdf</a> .
average.gene.stats (optional)	A flag to indicate whether to run the average gene statistics. This is a long process. Default: <i>don't run</i>
flip.strands (optional)	A flag to indicate whether the data is flipped, so a forward gene would appear as a reverse gene. Default: <i>don't flip</i>
output.prefix (optional)	A prefix for the output files. Default: <sam.file_basename>

## Input Files

1. sam.file  
Contains all the mapped reads for this library.

## Output Files

### MATLAB File:

1. <output.prefix>.mat: raw read coverage of each location in the genome for each strand.
2. <output.prefix>.allExpression.mat: the expression values calculated from this library for each gene.
3. <output.prefix>.zeroCov.mat: the tables used for generating a figure similar to Figure 5d in Levin et al.
4. <output.prefix>.segmentedGenes\_pooled.mat: the tables used for generating a figure similar to Figure 5a in Levin et al.

### Genome Browser Tracks:

5. <output.prefix>.norm.F.bedGraph: normalized forward track to be uploaded to a visualizer such as the [Integrative Genomics Viewer](#).
6. <output.prefix>.norm.R.bedGraph: normalized reverse track to be uploaded to a visualizer such as the [Integrative Genomics Viewer](#).

# GenePattern

7. <output.prefix>.F.bedGraph: non-normalized forward track to be uploaded to a visualizer such as the [Integrative Genomics Viewer](#).
8. <output.prefix>.R.bedGraph: non-normalized reverse track to be uploaded to a visualizer such as the [Integrative Genomics Viewer](#).

## Figures:

9. <output.prefix>.PooledVSlib.png: scatter plot of this library's expression profile and the pooled library from Levin et al.
10. <output.prefix>.PooledVSlib.qqplot.png: qq-plot of this library's expression profile and the pooled library from Levin et al.
11. <output.prefix>.dUTP\_controlVSlib.png: scatter plot of this library's expression profile and the dUTP\_control library from Levin et al.
12. <output.prefix>.dUTP\_controlVSlib.qqplot.png: qq-plot of this library's expression profile and the dUTP\_control library from Levin et al.
13. <output.prefix>.RNAvsDNAArraysVSlib.png: scatter plot of this library's expression profile and the RNA vs. DNA arrays as described in Levin et al.
14. <output.prefix>.RNAvsDNAArraysVSlib.qqplot.png: qq-plot of this library's expression profile and the RNA vs. DNA arrays as described in Levin et al.
15. <output.prefix>.zeroCoverage\_SUB\_pooledXaxis.png: a figure, corresponding to Fig 5c from Levin et al, using this library's data.
16. <output.prefix>.avgGene\_SUB.png: plot of average gene coverage.

## Expression files:

17. <output.prefix>.ExpressionProfile.tab: tab file with the expression profile of each gene in this library.
18. ExpressionProfile\_pooled.tab: tab file with the expression profile of each gene in the pooled library.
19. ExpressionProfile\_dUTP\_control.tab: tab file with the expression profile of each gene in the dUTP\_control library.

## Final Excel file:

20. <output.prefix>.final.xls: contains information on strand specificity, average CV, complexity, average number of segments per gene, and expression statistics.

## Example Data

To see the kind of output that you will receive with your own data, you can use [ftp://ftp.broadinstitute.org/pub/genepattern/rna\\_seq/StrandSpecRNASeqEval/dUTP.sam](ftp://ftp.broadinstitute.org/pub/genepattern/rna_seq/StrandSpecRNASeqEval/dUTP.sam) as input for an example run, with the flip.strands parameter set to on.

A sample MATLAB file that can be used to evaluate the execution of the pipeline is located here:

[ftp://ftp.broadinstitute.org/pub/genepattern/rna\\_seq/StrandSpecRNASeqEval/dUTP.76.pairs.mat](ftp://ftp.broadinstitute.org/pub/genepattern/rna_seq/StrandSpecRNASeqEval/dUTP.76.pairs.mat)

## Platform Dependencies

<b>Module type:</b>	RNA-seq
<b>CPU type:</b>	any
<b>OS:</b>	Linux
<b>Language:</b>	MATLAB